

7

Statistical Inference and Hypothesis Testing in Health Research

From Sample to Population

Statistical inference is the process of drawing defensible conclusions about a population from a sample, together with a rigorous statement of how confident that conclusion should be. Every quantity computed in Chapter 5's EDA — a mean, a proportion, a difference between groups — is an estimate, and this chapter develops the tools (standard errors, confidence intervals, hypothesis tests) needed to attach a defensible measure of uncertainty to that estimate.

Confidence Intervals and Hypothesis Tests

A confidence interval expresses a range of plausible values for a population parameter, given the observed sample; a 95% confidence interval, correctly interpreted, means that if the sampling and estimation procedure were repeated many times, approximately 95% of the resulting intervals would contain the true parameter value — it does not mean there is a 95% probability that this particular interval contains the true value. A hypothesis test formalises a comparison against a null hypothesis (typically 'no difference' or 'no association') and produces a p-value: the probability, if the null hypothesis were true, of observing a result

at least as extreme as the one obtained. A small p-value is evidence against the null hypothesis; it is not, by itself, a measure of clinical or public health importance, and this distinction between statistical significance and practical significance is one this chapter insists on throughout — a difference in mean birthweight of 15 grams between two groups can be statistically significant in a very large sample while being clinically negligible.

- Two-sample t-test / Mann-Whitney U test: comparing a continuous outcome (e.g., mean CD4 count) between two independent groups, using the non-parametric Mann-Whitney U when normality is doubtful or sample sizes are small.
- Paired t-test / Wilcoxon signed-rank test: comparing before/after measurements on the same subjects (e.g., pre- and post-treatment schistosomiasis egg counts).
- Chi-squared test / Fisher's exact test: testing association between two categorical variables (e.g., diagnostic test result by season), with Fisher's exact test preferred when expected cell counts are small.
- One-way ANOVA / Kruskal-Wallis test: comparing a continuous outcome across three or more groups (e.g., mean TPR across three ecological zones).

The Multiple Comparisons Problem

Testing many hypotheses in a single dataset — as happens routinely when comparing an outcome across dozens of LGAs, or screening many candidate risk factors for an outcome — inflates the overall (family-wise) chance of a false positive well beyond the nominal 5% used for any single test. Corrections such as the Bonferroni adjustment (conservative, dividing the significance threshold by the number of tests) or the Benjamini-Hochberg procedure (controlling the false discovery rate, generally preferred when many tests are run, as in genomic or wide screening studies) should

be applied whenever a study design involves testing more than a small, pre-specified number of hypotheses.

Survey-Weighted and Design-Based Inference

As introduced in Chapter 2, the NDHS and MICS use a complex, stratified, multi-stage cluster design. Ignoring this design — treating the sample as if it were a simple random sample — typically understates standard errors, because individuals within the same cluster (village or enumeration area) tend to be more similar to each other than to individuals in a different cluster (a phenomenon quantified by the intraclass correlation coefficient and summarised in the design effect). Correct analysis uses survey-specific software (R's 'survey' package or Python's 'samplix') that incorporates sampling weights, strata, and primary sampling units, producing standard errors that are typically wider — and therefore more honest — than a naive calculation.

Respondent-Driven Sampling: A Special Case

Respondent-driven sampling (RDS), used to reach populations without a sampling frame — such as female sex workers or other key populations relevant to HIV programming in Nigeria — recruits participants through peer referral chains rather than random selection, and standard survey-weighting methods do not directly apply. RDS-specific estimators (such as the Volz-Heckathorn estimator, successor-of-successor estimator, and RDS-II variants) adjust for each participant's network size and for homophily (the tendency of peers to recruit similar peers) to produce a population-representative estimate. Comparing multiple RDS estimators on the same dataset — a standard robustness check — often reveals that estimates are sensitive to the assumed recruitment model, which is why a careful analyst reports a range of estimators rather than a single point estimate for RDS-derived prevalence figures.

Worked Example

A Full Hypothesis Test, Start to Finish

Suppose we wish to test whether mean birthweight differs between infants of mothers who attended four or more ANC visits ($n = 180$, mean = 3120g, SD = 410g) and those who attended fewer than four ($n = 95$, mean = 2980g, SD = 450g). The null hypothesis is that the true population mean birthweight is equal in both groups. An independent two-sample t-test (Welch's version, which does not assume equal variances) gives a test statistic of approximately $t = 2.51$ with roughly 180 degrees of freedom, corresponding to a two-sided p-value of about 0.013. Because $0.013 < 0.05$, we reject the null hypothesis and conclude there is statistically significant evidence of a difference in mean birthweight between the two groups. The 95% confidence interval for the difference (140g) is approximately (30g, 250g). Note, following Section 7.2, that a 140g mean difference — while statistically significant and plausible given known physiology — should also be interpreted for its practical importance: is a 140g average difference large enough to change clinical or programmatic practice? That is a judgement the statistics alone cannot make.

R example

```
# R: full two-sample t-test workflow with confidence interval
anc_high <- rnorm(180, mean = 3120, sd = 410) # simulated for illustration
anc_low  <- rnorm(95, mean = 2980, sd = 450)

test_result <- t.test(anc_high, anc_low, var.equal = FALSE)
print(test_result)

cat("Mean difference:", round(diff(rev(test_result$estimate)), 1), "g\n")
cat("95% CI:", round(test_result$conf.int[1],1), "to", round(test_result$conf.int[2],1),
    "g\n")
```

Multiple Comparisons in Practice: A 20-LGA Example

Suppose an analyst tests whether DPT3 coverage differs from the national average in each of 20 LGAs, running 20 separate one-

sample tests at $\alpha = 0.05$. Even if there were truly no difference anywhere, we would expect roughly $20 \times 0.05 = 1$ false positive by chance alone. Applying a Bonferroni correction (Section 7.3) means each individual test must now achieve $p < 0.05/20 = 0.0025$ to be declared significant — a substantially stricter bar — while the Benjamini-Hochberg procedure instead ranks all 20 p-values and compares each to a sliding threshold, generally retaining more true positives than Bonferroni while still controlling the overall false discovery rate. The code below demonstrates both corrections on a simulated set of 20 p-values.

PYTHON example

```
# Python: applying Bonferroni and Benjamini-Hochberg corrections to 20 p-values
import numpy as np
from statsmodels.stats.multitest import multipletests

np.random.seed(1)
raw_pvalues = np.random.uniform(0, 1, 20)      # simulated p-values from 20 LGA tests
raw_pvalues[3] = 0.001                        # one genuinely strong signal

bonferroni_reject, bonferroni_p, _, _ = multipletests(raw_pvalues, alpha=0.05,
method="bonferroni")
bh_reject, bh_p, _, _ = multipletests(raw_pvalues, alpha=0.05, method="fdr_bh")

print("Bonferroni significant LGAs:", np.where(bonferroni_reject)[0])
print("Benjamini-Hochberg significant LGAs:", np.where(bh_reject)[0])
```

Worked Example

Survey-Weighted vs. Naive Standard Errors

Using a simulated NDHS-like dataset with 30 clusters of 20 households each, suppose the naive (unweighted, unclustered) standard error for estimated stunting prevalence is 1.2 percentage points, but the correct, design-based standard error (accounting for clustering and sample weights) is 2.1 percentage points — a design effect of roughly $(2.1/1.2)^2 \approx 3.1$, meaning the effective sample size is about one-third of the nominal sample size. A 95% confidence interval built on the naive standard error would be misleadingly narrow, potentially declaring a year-on-year change

in stunting prevalence ‘significant’ when the correctly-computed interval shows it is well within the range of sampling variability.

R example

```
# R: survey-weighted estimation using the survey package
library(survey)

ndhs_design <- svydesign(ids = ~cluster_id, strata = ~state,
                      weights = ~sample_weight, data = ndhs_data)

svymean(~stunted, ndhs_design)      # design-based mean and standard error
# Compare to the naive (incorrect) calculation:
mean(ndhs_data$stunted); sd(ndhs_data$stunted) / sqrt(nrow(ndhs_data))
```

Effect Sizes

Going Beyond Significance

Alongside a p-value, an effect size — a standardised measure of the magnitude of a difference or association, independent of sample size — allows results to be compared across studies and gives a more complete picture than significance alone. Cohen’s d (for a difference in means, expressed in standard deviation units) and the odds ratio or relative risk (for a difference in proportions) are the effect sizes most commonly reported in Nigerian health research; conventional (though somewhat arbitrary) benchmarks describe a Cohen’s d of 0.2 as small, 0.5 as medium, and 0.8 as large, but the appropriate benchmark for ‘clinically meaningful’ should ultimately be set by domain knowledge of the outcome in question, not by these generic statistical conventions alone.

Extended Case Study

A Non-Significant Result That Was the Right Answer

This chapter has repeatedly emphasised that statistical significance is not the same as practical importance; it is equally important to internalise the reverse lesson, that a properly conducted, non-significant result can itself be the correct and most useful answer to a research question, rather than a disappointing failure to be

explained away. Consider a study testing whether a low-cost, community health worker-delivered adherence counselling intervention improves viral suppression rates among a cohort of 240 adults on antiretroviral therapy, compared to a control group of 235 receiving standard care. The observed difference in 12-month viral suppression is small (74% intervention vs. 71% control) and the associated 95% confidence interval for the difference (-4 to +10 percentage points) comfortably includes zero, giving a non-significant result ($p = 0.29$) by any conventional threshold.

A team under pressure to report a 'positive' finding might be tempted to search for a subgroup in which the effect appears larger (a form of the multiple-comparisons problem discussed in Section 7.3, sometimes called 'p-hacking' when done without pre-specification and full disclosure), or to describe the result using softened language such as 'a promising trend toward improved suppression' that overstates the actual evidence. The methodologically correct response, and the one this book argues for throughout, is instead to report the result plainly: this study did not find statistically significant evidence that the counselling intervention improved viral suppression, the confidence interval is consistent with anything from a modest harm to a moderately large benefit, and a substantially larger sample size would be needed to distinguish a true small-to-moderate effect from no effect at all. This honestly-reported null result remains genuinely valuable to the HIV programme deciding whether to scale up the intervention nationally: it correctly tells them that the evidence so far does not justify the additional cost of scale-up, while leaving open (via the confidence interval, not via speculative subgroup mining) the possibility that a moderate true benefit exists and a larger, adequately powered follow-up study would be needed to detect it with confidence.

One-Sided Versus Two-Sided Tests

This chapter's hypothesis tests have all been presented as two-sided, testing for a difference in either direction, which is the appropriate and more conservative default for most health research questions. A one-sided test, which tests only for a difference in one pre-specified direction (for example, testing only whether a new bednet reduces, rather than testing for any difference in either direction from, malaria incidence), produces a smaller p-value for the same data and is therefore easier to declare significant — a property that makes one-sided tests appropriate only when a difference in the opposite direction would be scientifically implausible or practically irrelevant, and when this choice is made and stated BEFORE seeing the data, never chosen after the fact simply because it produces a more favourable p-value for a result already observed.

Sample Size and Power Calculations Before Data Collection

This chapter's methods for interpreting a completed study's results are only half of the statistical inference picture; before data collection begins, a power calculation determines the sample size needed to have a reasonable chance (conventionally 80% or 90%) of detecting a specified minimum effect size, if that effect genuinely exists, at a chosen significance level. Skipping this step is a common and consequential planning error: a study that turns out to be underpowered for the effect size it was actually looking for will produce the kind of wide, inconclusive confidence interval discussed in Section 7.10's HIV counselling example — not because the intervention necessarily has no effect, but because the study was never designed with enough participants to distinguish a real, moderate effect from no effect at all. Performing a power calculation at the design stage, using a realistic, ideally literature-informed estimate of the minimum clinically important effect

size, is therefore a critical planning step that determines whether a study's eventual result — significant or not — will actually be informative.

Equivalence Testing

When 'No Significant Difference' Isn't Enough

A standard hypothesis test, as developed in this chapter, is designed to detect a difference from a null value; failing to reject the null hypothesis is not the same as positive evidence that two groups are truly equivalent, since a study may simply lack the statistical power to detect a real but modest difference (Section 7.12). When a research question genuinely requires evidence of equivalence — for example, showing that a cheaper, locally-produced diagnostic reagent performs no worse than an imported reference reagent, within a pre-specified clinically acceptable margin — equivalence testing (such as the two one-sided tests, or TOST, procedure) should be used instead of a standard test, since it is explicitly designed to provide positive statistical evidence of similarity within a defined margin, rather than merely failing to find a difference.

Bayesian Alternatives to Classical Hypothesis Testing

A Brief Orientation

Every hypothesis test presented in this chapter follows the classical (frequentist) framework, interpreting probability as long-run frequency across repeated sampling. The Bayesian alternative instead treats probability as a degree of belief, updated from a prior distribution to a posterior distribution as data accumulates, and produces a credible interval that can be interpreted directly as 'there is a 95% probability the true parameter lies in this range' — the intuitive interpretation many practitioners mistakenly apply to a classical confidence interval, but which is only strictly correct in the Bayesian framework. Bayesian methods are particularly well

suited to situations with genuinely informative prior knowledge (for example, incorporating a well-established prior estimate of malaria RDT sensitivity from a large prior meta-analysis when analysing a small new local validation study) and to sequential decision-making contexts (such as an outbreak response) where beliefs need to be continuously updated as new data arrives, though a full treatment of Bayesian inference is beyond this book's scope.

A Reusable, Assumption-Checking Hypothesis Test Function

The worked hypothesis test in Section 7.6 assumes the analyst will remember to check normality and variance-equality assumptions manually before choosing between a t-test and its non-parametric alternative. A more robust function automates this check, selecting the appropriate test itself and reporting which test it used and why — reducing the chance that an assumption check is quietly skipped under time pressure, while still keeping the choice fully transparent and auditable in the function's output.

R example

```
# R: an assumption-checking wrapper that selects t-test or Mann-Whitney automatically
compare_two_groups <- function(x, y, alpha = 0.05) {
  norm_x <- shapiro.test(x)$p.value > alpha
  norm_y <- shapiro.test(y)$p.value > alpha
  if (norm_x && norm_y) {
    test <- t.test(x, y)
    method <- "Welch two-sample t-test (normality not rejected in either group)"
  } else {
    test <- wilcox.test(x, y)
    method <- "Mann-Whitney U test (normality rejected in at least one group)"
  }
  list(method = method, statistic = test$statistic, p_value = test$p.value)
}

result <- compare_two_groups(anc_high, anc_low)
print(result$method) # states which test was used and why, not just the p-value
```

Chapter Summary

- A confidence interval and a p-value both quantify sampling uncertainty but answer different questions; neither replaces judgement about clinical or public health importance.
- The correct test depends on data type (continuous/categorical), number of groups, and paired/unpaired structure.
- Testing many hypotheses without correction inflates the false positive rate; Bonferroni and Benjamini-Hochberg are standard corrections, with BH generally retaining more power.
- Complex survey data (NDHS/MICS) requires design-based (weighted, clustered) inference, not naive simple-random-sample formulas; ignoring design can understate standard errors by a factor of two or more.
- Respondent-driven sampling requires specialised estimators, and results should be reported across multiple RDS estimators to assess robustness.
- A complete hypothesis test reports the test statistic, p-value, confidence interval, and an explicit judgement of practical significance — not the p-value alone.

Mini-Project

Mini-Project 7: A Multi-LGA Hypothesis Testing Report with Correction

Conduct hypothesis tests comparing an indicator across at least 10 groups (LGAs, facilities, or states), correctly applying a multiple comparisons correction.

Step 1. Obtain or simulate an indicator (e.g., mean birthweight, DPT3 coverage) for at least 10 groups, each compared against a fixed benchmark (e.g., the national average).

Step 2. Run an appropriate hypothesis test (Section 7.2) for each group without correction, and report how many are 'significant' at $\alpha = 0.05$.

Step 3. Apply both a Bonferroni and a Benjamini-Hochberg correction (Section 7.7) and report how the number of significant results changes under each.

Step 4. For at least one significant result, report the effect size and confidence interval, and comment on its practical (not just statistical) significance.

Step 5. Write a one-paragraph recommendation for a programme manager, explaining in plain language why the corrected results are more trustworthy than the uncorrected ones.

Deliverable: A short report containing the uncorrected and corrected significance counts, one full worked hypothesis test with effect size and CI, and the plain-language recommendation.

End-of-Chapter Exercises

1. A study reports $p = 0.03$ for a mean difference of 0.2 g/dL in haemoglobin between two groups of 5,000 participants each. Discuss statistical versus clinical significance.
2. You test 20 LGAs for a difference in vaccination coverage against a national benchmark. Explain why a Bonferroni or Benjamini-Hochberg correction is appropriate, and compute the Bonferroni-adjusted threshold.
3. Explain, without formulas, why ignoring cluster design in an NDHS analysis tends to produce overly narrow confidence intervals.
4. Using the worked example in Section 7.6, recompute (by hand or in code) the approximate 95% confidence interval for the birthweight difference if the sample sizes had instead been 60 and 40 (all other figures unchanged), and comment on how the interval width changes.
5. In Section 7.7, explain in your own words why the Benjamini-Hochberg procedure typically flags more LGAs as significant than the Bonferroni correction, for the same set of p-values.
6. Using the R code in Section 7.8, explain what a design effect of 3.1 means in practical terms for the effective sample size of a 600-household survey.
7. Two RDS estimators applied to the same key-population dataset give HIV prevalence estimates of 14% and 19%. Explain why this discrepancy is not a ‘mistake’ and how you would report both figures responsibly.
8. Calculate Cohen’s d for the birthweight example in Section 7.6 (mean difference 140g, pooled SD approximately 425g), and classify it as small, medium, or large using the conventional benchmarks in Section 7.9.

9. Explain why an effect size is more comparable across different studies than a p-value alone.
10. Why might the conventional ‘small/medium/large’ benchmarks for Cohen’s d be inappropriate for judging whether a specific clinical finding is meaningful?
11. In the Section 7.10 case study, explain why searching for a subgroup with a larger apparent effect, after seeing the main non-significant result, would have been methodologically inappropriate without pre-specification.
12. Using the reported 95% CI (-4 to +10 percentage points) in Section 7.10, explain what range of true effects (harmful, none, or beneficial) remains consistent with this study’s data.
13. Why is ‘a promising trend toward improved suppression’ considered an overstatement of the evidence described in Section 7.10?
14. What would a well-justified next step be for the HIV programme, given the non-significant result and wide confidence interval in Section 7.10?
15. Explain briefly why a larger sample size, rather than a different statistical test, is generally the correct remedy for the kind of inconclusive result described in Section 7.10.
16. Using Section 7.11, explain why choosing a one-sided test AFTER seeing the data would be methodologically inappropriate, even if the direction chosen matches the data.
17. Give an example of a Nigerian health research question where a one-sided test, pre-specified before data collection, would be scientifically justified.
18. Why does a one-sided test produce a smaller p-value than a two-sided test for the same data and same direction of effect?

19. Explain why two-sided tests are described as the ‘more conservative default’ for most health research questions.
20. If a colleague proposes switching from a two-sided to a one-sided test partway through analysing already-collected data, what would you advise, and why?
21. Using Section 7.12, explain why the HIV counselling study in Section 7.10 might have been designed with too small a sample size for its true effect size.
22. What does it mean for a study to have ‘80% power’ to detect a specified effect size, in your own words?
23. Why is skipping a power calculation described as a ‘planning error’ rather than simply a matter of limited resources?
24. Propose how you would estimate a realistic minimum clinically important effect size before conducting a power calculation for a new study.
25. Explain the connection between an underpowered study and the wide, inconclusive confidence interval described in Section 7.10.
26. Summarise, in your own words, why this chapter treats statistical significance and practical importance as two distinct questions that must both be addressed.
27. Explain the common thread linking the multiple comparisons discussion (Section 7.3) and the power calculation discussion (Section 7.12).
28. Why does this chapter emphasise reporting confidence intervals and effect sizes, not just p-values, in every worked example?
29. Propose a research question from your own field where a poorly designed study (e.g., underpowered, or with an inappropriate one-sided test) could produce a misleading conclusion.

30. If you had to explain to a policymaker why a ‘non-significant’ result can still be useful, what would you say?
31. Using Section 7.14, explain in your own words the key conceptual difference between a classical confidence interval and a Bayesian credible interval.
32. Why might a Bayesian approach be particularly well suited to an active outbreak response context?
33. Describe a scenario where strong, well-established prior knowledge could reasonably inform a Bayesian analysis of a small new dataset.
34. What is the common (but technically incorrect, under the classical framework) way people often misinterpret a 95% confidence interval, and how does the Bayesian credible interval avoid this issue?
35. Would you consider a Bayesian or classical approach more appropriate for a large, well-powered national survey with no strong prior information? Explain your reasoning.